DOCUMENT RESUME

ED 360 326 TM 020 123

AUTHOR Jaeger, Richard M.; And Others
TITLE Integrating Multi-Dimensional Performances and Setting Performance Standards.
PUB DATE Apr 93
NOTE 27p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Classification; Comparative Analysis; Elementary School Teachers; Elementary Secondary Education; *Evaluators; Higher Education; *Performance; Profiles; Secondary School Teachers; Standards; Summative Evaluation; Teacher Certification; *Teacher Evaluation; Teacher Qualifications
IDENTIFIERS Angoff Methods; National Board for Professional Teaching Standards; *Performance Based Evaluation; Policy Capturing Method; *Standard Setting; Standards for Educational and Psychological Tests; Teacher Candidates

ABSTRACT

Methods that might be used to establish standards of performance that will permit the National Board for Professional Teaching Standards to classify candidate teachers as highly accomplished (worthy of National Board certification) or less than highly accomplished (not worthy of certification) are contrasted. Plans for research on these issues are discussed. Two approaches are considered. One, "policy capturing," is a judgmental process that attempts to elicit and characterize the decision strategies used by expert judges when they evaluate profiles of performances and reach a summative decision. The other method is an extension of the standard-setting method of W. H. Angoff (1971), in which judges are asked to estimate for individual test items the proportion of minimally qualified examinees who would answer the item correctly. Each method is examined in some detail, and the proposed strategies for applying them are explored. If carefully applied, these strategies should satisfy the requirements of the 1985 "Standards for Educational and Psychological Tests." One chart and four figures illustrate the discussion. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED360326

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
RICHARD M. JAEGER
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Integrating Multi-Dimensional Performances and Setting Performance Standards

Richard M. Jaeger
University of North Carolina, Greensboro

Barbara S. Plake
University of Nebraska-Lincoln

Ronald K. Hambleton
University of Massachusetts at Amherst

Presented at the annual meeting of the National Council on Measurement in Education in a symposium titled Advances in the Psychometrics of Performance Assessment: Old Problems – New Strategies for the National Board for Professional Teaching Standards, Atlanta, Georgia April 13-15, 1993.

TM020423

BEST COPY AVAILABLE

Standard-setting on certification examinations is time-consuming, expensive, and controversial. The methods used to establish performance standards become a central area of concern when questions about the validity of classification decisions arise. Although many well-known methods are available for setting standards (see Berk (1986) and Jaeger (1989) for reviews), most methods must be adapted to the situation in which they are used. The complex demands of a performance assessment situation stimulated the formulation of new standard-setting methodology that extends and complements the methods reviewed and summarized by Berk and Jaeger. That new methodology, and plans for examining its properties, are described here.

The impetus for this paper is the intention of the National Board for Professional Teaching Standards (NBPTS) to develop and operate a nationwide, voluntary program for assessing and certifying "highly accomplished" classroom teachers. Ultimately, the National Board intends to offer teacher certification in some 33 fields defined by subject-matter specialty and age range of students taught. At present, assessment packages are under development in the Early Adolescence English/Language Arts certification field and the Early Adolescence Generalist certification field. The Early Adolescence age range includes students who would normally be enrolled in grades 5 through 9, were they enrolled in graded schools. Generalist teachers are those who routinely teach several subject areas, such as mathematics and science or English and social studies.

Performance assessment is the method selected by the NBPTS to distinguish between teachers qualified to receive National Board Certification and those of lesser accomplishment. The National Board's assessment packages currently under

BEST COPY AVAILABLE

development contain a substantial number of performance exercises that will be completed by all candidates for certification. One set of these exercises will be completed by candidates at their school site. Another set will be completed in two days at an assessment center under standardized conditions. The school-site exercises might include, among others, such activities as collection and interpretation of artifacts of students' performances over a prescribed period of time, preparation of a videotape of teaching performance under prescribed conditions, collection and description of artifacts of instruction associated with a given instructional unit, and writing a reflective essay on the nature of various elements of a teaching portfolio and their relationships. The assessment center exercises might include, among others, such activities as preparation of a plan for an instructional unit using prescribed resources, analysis and critique of the teaching of a hypothetical colleague, evaluation and critique of a body of student work followed by prescription of appropriate remedial instruction, and development of a response to a simulated classroom management crisis.

Candidates' performances in response to each of these exercises will be evaluated using highly structured scoring rubrics that produce multiple scores with respect to a number of content standards of teacher accomplishment. The measurement problems are thus very challenging since candidates will engage in complex exercises that yield somewhat incommensurate multidimensional scores. Each exercise will yield a score on several content standards, and different exercises will yield scores on partially overlapping sets of content standards.

Establishing performance standards for the assessment packages of the National Board for Professional Teaching Standards will be substantially more challenging than setting standards on a competency test. The complexity of the task can be likened to judging the outcomes of a women's Olympic gymnastics

BEST COPY AVAILABLE

competition. In that assessment of athletic prowess, multiple exercises are included — floor exercises, uneven bars, balance beam, and the horse. Each exercise contributes, in varying degree, to an assessment of multiple content standards — style, originality, difficulty, form, gracefulness, and technical precision. Analogously, in the National Board's assessment, we have content standards developed by standards committees such as "Highly accomplished generalists regularly analyze, evaluate, and strengthen the effectiveness and quality of their practice." Just as gracefulness is assessed through several exercises in the Olympics — the uneven bars, the balance beam, and floor exercises — so too, in the National Board's assessment, each of the standards' committees' content standards will be assessed through candidates' performances on a number of exercises, such as portfolios, interviews, and simulations.

Multiple assessors will likely evaluate participants' levels of performance on each of exercises on each of a number of content standards, so as to increase the precision of overall judgments and reduce assessment bias. This is done in the Olympics and will be done in National Board assessments as well.

Of course, in the Olympics, the final outcome of assessment is not a certificate and candidates are judged against each other. In contrast, in the National Board's assessment, candidates are judged against well-specified, content-based standards, and at least in theory, every candidate could be a winner. Nonetheless, each assessment requires a well-specified, highly structured procedure for establishing standards of performance. And if final judgments are to be valid and reliable, judges must be highly skilled, well informed, and well trained. Since, as noted earlier, all of the National Board's assessment exercises are performance-oriented, almost none of the well-researched, popular methods for setting performance standards are directly applicable (e.g., the Angoff, 1971; Ebel, 1972; Nedelsky, 1954; or

Jaeger, 1982 methods). The methods that might be adapted readily to performance assessments, such as the contrasting groups approach (Livingston & Zieky, 1982), require the identification of highly-accomplished and less-than-highly-accomplished teachers at the outset. If such distinctions could be make with confidence, the assessment exercises and, indeed, the National Board's assessment development project, would not be needed.

Performance standards must be established for two levels of assessment — one at the content-standard or exercise level, and the other for candidates' performances on an entire assessment package. This need is unusual, and the literature does not describe the application or adaptation of any widely used standard-setting procedure to a situation such as this. The obvious use of conventional standard-setting methods, for establishing individual passing scores on each exercise/content-standard combination, with the subsequent adoption of a "multiple hurdles model" for setting overall performance standards, is neither practical nor appealing for National Board Certification. The exercises that contribute to the assessment of competence on each content standard will vary substantially in their importance, and probably in their reliability and in the variability of the performances they elicit as well. Therefore, weightings that reflect the judged importance of candidates' performances on the various exercises that compose a National Board assessment package must be elicited and made a part of the standard-setting process. This too is uncommon. The aggregation of scores for various exercise/content-standard combinations to obtain performance standards for each content standard will not be as simple as summing over items in paper-and-pencil tests to obtain total scores. Profiles of performance across exercise/content-standard combinations may be more important than any sum of scores.

BEST COPY AVAILABLE

The purpose of this paper is to describe and contrast methods that might be used to establish standards of performance that will permit the National Board for Professional Teaching Standards to classify candidate teachers as "highly accomplished," and therefore worthy of National Board Certification, or "less than highly accomplished," and therefore not qualified to receive National Board Certification. Although the standard-setting methods described in this paper were conceived to address the needs of the National Board for Professional Teaching Standards, they are clearly applicable in any assessment context that requires integration of performance assessment results across multiple sources of evidence to yield a pass/fail decision.

A Summary of Standard-Setting Approaches to be Examined

Two approaches to setting performance standards on the National Board's assessment packages will be examined. One, termed "policy capturing," is a judgmental process that attempts to elicit and characterize the decision strategies employed by expert judges when they evaluate profiles of candidates' performances on multiple exercises and reach a summative decision concerning the overall quality of the candidates' performances. In this process, expert judges respond independently to a large number of simulated profiles of candidates' performances on the elements of an assessment package, indicating for each stimulus profile, their evaluation of the overall performance of the candidate. These profile-response pairs are then used to "capture a judge's policy" in evaluating overall candidate performance. Various analytic procedures are applied to the data contained in the profile-response pairs to determine the relative weights judges apply to elements of candidate performance in reaching an overall evaluation, and the range of performance profiles associated with a recommendation to "pass" (in this case,

BEST COPY AVAILABLE

certify) a candidate. The method produces a mathematical representation of the judgment policy of each member of a judgment panel, an estimate of the decision consistency of each panel member, and information on the inter-judge consistencies of panel members.

The second method to be explored is an extension of the well-known method for setting standards on pencil-and-paper tests that was proposed by William Angoff in 1971 (although he has attributed the idea to Ledyard Tucker). As Angoff's method is conventionally applied, judges are presented with individual test items and are asked to estimate for each item, the proportion of minimally qualified examinees who would be able to answer the item correctly. In the performance assessment context of the National Board, as in many others, a panel of expert judges would work independently. They would then be provided an opportunity to reconsider their initial judgments following a controlled discussion during which they would justify their initial recommendations. The judgment task would be to consider each exercise in an assessment package and estimate the score that would be earned by a candidate who barely satisfied the requirements for National Board Certification as a highly accomplished teacher. Judges would also be asked to assign an importance rating to each exercise, in determining candidates' eligibility for National Board Certification. The recommendations of individual judges would be aggregated to determine a performance standard for each exercise with respect to each content standard, and a rule for weighting candidates' performances in determining their overall performance on the content standard. They would also be asked to provide importance ratings for content standards that then could be aggregated to determine an overall performance decision on an assessment package.

The next section of this paper contains an elaborated description of these methods of setting performance standards, together with a description of our

BEST COPY AVAILABLE

intended strategy for applying the methods. Alternative conceptual approaches to the integration of multidimensional performance profiles, including compensatory models, conjunctive models, and disjunctive models are among the issues discussed.

A Policy-Capturing Approach to Standard Setting: The Method and Our Proposed Application

We wish to know the salient information used by judges when they determine whether a candidate's performances on the set of exercises that assess a given content standard are sufficient to warrant the classification of that candidate as "highly accomplished" on that content standard. One way to do this is to analyze judges' classifications of hypothetical patterns of candidates' performances on the set of assessment exercises. As noted earlier, in a policy capturing procedure, judges are presented with a large number of profiles (such as the one illustrated in Figure 1) of the performances of hypothetical candidates for certification. Working

---

Insert Figure 1 Here

---

independently, judges are asked to make a summative judgment of the quality of the overall performance of each hypothetical candidate, based on the performance profile of that candidate. The set of profiles, together with the judges' ratings of overall performance, are then used to "capture the judges' policies" in awarding overall performance ratings. This procedure yields the following information:

1. Weights for the respective assessment exercises that could be used to classify candidates into performance categories on each content standard;

2. An indication of the consistency of ratings produced by individual judges;

BEST COPY AVAILABLE

3. Individual performance standards for each of the content standards that could be used in eliciting, or in developing methods for eliciting, overall performance standards for an entire assessment package.

The policy capturing procedure has been used in the field of industrial and organizational psychology, where it has been applied to personnel selection problems (cf., Hobson, Mendel & Gibson, 1981; Hobson & Gibson, 1983; Stumpf & London (1981)) and in the field of urban planning, where it has been applied to problems of site selection for public service agencies (cf., Gardiner & Edwards (1975)). These problems present superficially similar decision contexts. In each, a selection decision must be made. In each, the decision must be based on a large number of factors that define each alternative (In the case of personnel selection, such factors as education, prior work experience, salary demands, performance test scores, interviewers' impressions, etc. In the case of site selection for a governmental facility, such factors as ease of access, cost of land, proximity to residential neighborhoods, availability of public transportation, availability of parking, etc.). The similarity of the decision situation facing the National Board for Professional Teaching Standards is obvious. The Board will have to consider a candidate's performances on a wide range of assessment exercises in making a decision to award or withhold certification.

In the application of policy capturing we propose to explore, the procedure will be used at two levels. First, performance judgments from the smallest scoreable units that result from a National Board assessment of teachers (content-standard scores by exercise) will be averaged across assessor-assigned scores at the scoreable-unit level. Then these averages will be combined by using the weights that result from application of a policy capturing procedure to produce performance standards at the content-standard level. Second, policy capturing will be applied to candidates'

BEST COPY AVAILABLE

performance scores at the content-standard level to determine weights that can be used to place candidates into "certify" and "do-not-certify" categories.

Keep in mind that the NBPTS assessment packages are composed of multiple exercises that measure multiple content standards, and that each exercise is to be scored by multiple assessors. Each exercise will thus yield a number of scores for each candidate. In tabular form, the pattern of a candidate's assessment scores might look like the table on the next page. The entries in this table, designated by the symbols A1, A2, ... , A10 denote assessors who will judge the quality of candidates' performances on certain National Board exercises in relation to certain content standards. Note that different pairs of assessors contribute to an individual candidate's performance score on any content standard. Note also, that in at least some cases, different assessors judge candidates' performances on aspects of a given exercise that provide information for scoring different content standards. In the table that follows, we have used hypothetical labels for exercises and content standards that might reflect a portion of a National Board for Professional Teaching Standards assessment package.

The first proposed application of policy capturing would yield weights for aggregating a candidate's scores on exercise-by-content-standard combinations in a particular row of the table, to compute a performance score for the content standard that identified the row. Separate policy capturing procedures that employed distinct panels of judges would be used to obtain weights for aggregating scores in each row of the table. The second application of policy capturing would yield a set of weights for aggregating candidates' performance scores across content standards (rows of the table) to yield an overall "certify/do-not-certify" decision.

BEST COPY AVAILABLE

	Exercise 1 Prepare Plan for Instruction	Exercise 2 Video of Teaching and Essay on it	Exercise 3 Simulated Class-room Crisis	Exercise 4 Analysis of Other's Teaching	Exercise 5 Artifacts of Student Perform-ances	Exercise 6 Critique Student Work	Exercise 7 Reflect-ive Essay on Teaching
Standard 1 (e.g., Teacher is committed to students)	A1 A2		A1 A2		A3 A4		
Standard 2 (e.g., Teacher knows the subject they teach and how to teach it)					A3 A4	A3 A4	A4 A5
Standard 3 (e.g., Teacher manages and monitors student learning)	A1 A2	A5 A6			A7 A8		
Standard 4 (e.g., Teacher thinks system-atically about teaching)		A5 A6		A5A6	A9 A10	A9 A10	A9 A10

In our exploration of policy capturing, we intend to use about 15 panels of judges (one for each content standard) since our "table" will have about 15 rows. Each panel will consist of 10 members. The panels will be selected from the population of teachers who will be used as assessors by the National Board for Professional Teaching Standards in large-scale field tests of its assessment packages during the 1993-94 school year. To the extent possible, assessors used in this study will be selected so as to represent the entire population of assessors used in the 1993-94 field tests, in terms of geographic region, gender, age, and subspecialty.

Each member of a judgment panel will be trained to assess candidates' performances on the exercises that provide information for assessing candidates'

BEST COPY AVAILABLE

performances on a particular content standard. Two alternative standard-setting methods (policy capturing and an extended Angoff method, described below) will be applied by each panel of judges; administration of the methods will be in counter-balanced order to control for potential order effects.

For each content standard, out of all possible exercise score profiles for hypothetical candidates for certification (with k score points on each of a exercises, a total of k^a profiles exist), 200 profiles will be selected randomly for study. The configuration of selected score profiles will mirror the covariance structure of scores observed in small-scale field tests conducted by the assessment package developers. In addition to these 200 profiles, 10 profiles will be repeated for the purpose of assessing judges' rating consistency.

After being trained to use policy capturing, each panel member will consider the set of 210 score profiles independently. In response to each profile, each panel member will provide a judgment of the performance of the hypothetical candidate on the content standard assessed. This criterion variable will be assigned a score on a scale with a range from one to four as follows: 1 = Novice, 2 = Journeyman, 3 = Accomplished, and 4 = Highly Accomplished. Careful attention will be given to the preparation of definitions of the score points on the criterion variable.

These data will be used to estimate the policy employed by each panel member in weighting the importance of the exercises that assess the content standard they are judging. Panel members will also be asked to complete an instrument designed to assess their comfort with the process and their confidence in their ratings.

BEST COPY AVAILABLE

Several methodological issues will be studied:

1. The comparability of policy capturing results for randomly selected subsets of 5 persons drawn from the 10 panel members,

2. Differences in the stability of policy capturing results produced from randomly selected subsets of 100 profiles from the set of 200, and from non-representative samples such as those that might be found at the higher end of the score scale, and

3. The potential of using a dichotomous criterion variable; i.e., a variable that is scaled as "highly accomplished" versus "less-than-highly accomplished."

A variety of analytic models will be applied to the judgment data that are provided by the policy capturing procedure: (a) ordinary least squares estimation of the parameters of compensatory, conjunctive, and disjunctive models; (b) logistic regression, and (c) combination models involving, for example, absolute thresholds and compensatory models. Compensatory, conjunctive, disjunctive, and combination models are defined as follows:

A *Compensatory Model* assumes that highly accomplished performance on one content standard can compensate for less accomplished performance on another when the overall performance of a candidate is judged. For example, highly accomplished performance in classroom management can compensate for less accomplished performance in instructional planning in seeking National Board Certification. An illustration of the hyper-plane-like response surface that results from the application of this model is illustrated in Figure 2. The illustration assumes that candidates are evaluated on only two certification standards and that an overall evaluation of the

BEST COPY AVAILABLE

candidate's performance is provided on a four-point scale, such as the one described earlier.

Insert Figure 2 Here

A *Conjunctive Model* requires that a candidate be highly accomplished on all content standards if (s)he is to be considered a highly accomplished teacher. A candidate would be required to exhibit highly accomplished performance on all National Board standards to gain National Board Certification. An illustration of the hyper-paraboloid-like response surface that results from application of this model is illustrated in Figure 3. The assumptions underlying this figure are identical to those underlying the illustration in Figure 2.

Insert Figure 3 Here

A *Disjunctive Model* requires that a candidate be highly accomplished on only one or two content standards in order to be granted National Board Certification. For example, highly accomplished performance in knowledge of student development and instructional management might suffice. An illustration of the hyper-hyperboloid-like response surface that results from application of this model is illustrated in Figure 4. Again, the assumptions underlying this figure are identical to those underlying the illustration in Figure 2.

Insert Figure 4 Here

BEST COPY AVAILABLE

A *Combination Model* might require a candidate to achieve specified standards of performance on all content standards, but to exceed those performance standards by exhibiting exceptional performance on several of the content standards. In this model, achievement of specified minimum performance levels on all content standards would not be sufficient to earn National Board Certification. For example, candidates might have to exhibit accomplished practice on all content standards, and exhibit highly accomplished practice on a subset of them.

The policy capturing procedure requires the use of judges who are intimately familiar with the assessment dimensions being applied in the decision situation in which it is used. In the case of National Board assessments, judges would have to be knowledgeable about exercises in the Board's assessment packages and about the procedures and criteria used to score candidates' performances on those exercises. In addition, judges would have to be knowledgeable about the content standards adopted by the National Board's standards committees and about the Board's definitions of high levels of teacher accomplishment.

The policy capturing procedure produces a judgment policy for each member of a judgment panel. These policies must be aggregated across judges to create an overall policy. Several alternative methods of aggregation will be examined in this study. A preliminary cluster analysis of panel members' recommended policies will be conducted to determine how many clusters of policies exist. If a single-cluster solution appears, panel members' policies will be averaged, or the median values of panel members' policies will be assumed to represent the entire cluster. If several clusters are found, information on panel members' backgrounds will be examined to explore the basis of the clusterings. Some clusters of judges might be regarded as having the greatest expertise, and their recommendations might then be averaged.

BEST COPY AVAILABLE

As a check on the cluster analysis results, correlations between profiles of weights will be computed for each pair of panel members, and these profiles will be treated as proximities in a non-metric multidimensional scaling analysis. The one- and two-dimensional multidimensional scaling solutions will be represented graphically, as will two-dimensional projections of higher-dimensional solutions, to yield a visual confirmation of the clustering of panel members' recommended profiles.

Application of the policy capturing procedure in the study we envision can be summarized as follows:

1. Panels of expert judges will be selected.

2. Panel members will complete the assessment package.

3. Panel members will be trained in scoring the assessment exercises.

4. Panel members will be given hypothetical profiles of performance on the components of the assessment package, and asked to assign corresponding evaluative scores at the level of content standards, and, thereafter, at the level of the overall assessment package.

5. Resulting data will be analyzed to capture the scoring "policy" of each panel member.

6. Panel members' consistencies will be estimated.

7. The judgments of consistent panel members will be aggregated to produce an overall "policy" and to elicit resulting performance standards.

8. Judgment data elicited from panel members will be analyzed using a variety of models that reflect alternative aggregation rules.

9. The profiles of weights recommended by individual judges will be cluster analyzed and will be used in a multidimensional scaling analysis to determine whether simple aggregation of recommendations across judges would adequately represent an entire panel.

BEST COPY AVAILABLE

An Extended Angoff Approach to Standard Setting: The Method and Our Proposed Application

The extended Angoff method of standard setting is an adaptation of a procedure that has been used extensively to set standards on certification tests and competency tests. It would operate as follows: Based on their conceptualization of a "just highly accomplished candidate," a panel of judges convened to produce a performance standard for a given content standard would individually and independently determine, for each exercise that assessed that content standard, the expected score on each scoring rubric that would be earned by a randomly selected, hypothetical, "just qualified" candidate. A performance standard would be computed for each exercise/content-standard combination by averaging the recommendations provided by each judge. These recommended standards would then be used in computing a weighted average across all exercises that assessed a given content standard. The weights used would be a function of (a) a consensus rating (provided by the panel of judges) of the importance of the various exercises in assessing candidates' performances on the content standard, (b) the variability of observed scores on the various exercises, and (c) the reliabilities of the scores produced by the various exercises for use in assessing candidates' performances on the content standard.

Aggregation of these weighted scores across the assessment exercises used to assess each content standard would yield:

1. Classification of candidates into performance categories on each content standard;
2. Individual performance standards for each of the content standards that could be used in eliciting, or developing methods for eliciting, overall performance standards for the entire assessment package.

BEST COPY AVAILABLE

Panel members who apply the policy capturing procedure also will be trained to apply the extended Angoff standard-setting method. They also will be guided in conceptualizing a hypothetical "just-highly-accomplished" teacher. As specified earlier, panel members will individually and independently specify the score they would expect a "just-highly-accomplished" teacher to earn on each of the exercises associated with the content standard they are considering. In addition, the panel members will be asked to indicate how important they believe each exercise is in contributing to an assessment of a candidate's performance on the content standard they are considering. After providing their initial estimates, panel members will be provided with the following relevant data: 1) Performance standards recommended by the other panel members, and 2) Importance ratings provided by the other panel members. The panel members will be given an opportunity to discuss this information prior to providing their final estimates of expected scores and importance ratings for each exercise that assesses the content standard under consideration. Panel members will also be asked to complete an instrument designed to assess their comfort with the process and their confidence in the quality of their ratings.

Specific features of this method that will be studied are: (a) the phrasing of the stimulus questions presented to the panel members, (b) the desirability of having panel members discuss their initial judgments and their subsequent judgments, (c) the utility of a discussion and solicitation of recommendations by panel members of procedures for aggregating weighted scores within content standards. These and other features of the method (such as the definition of a "just-highly-accomplished" teacher, the nature and amount of training provided to panel members, the design of rating forms) also will be investigated in a pilot study.

BEST COPY AVAILABLE

Data will be analyzed as follows: For each exercise that provides information for assessing a content standard, recommended scores will be averaged across panel members. These averaged scores will be used to compute a weighted average across all exercises that provide information for assessing a given content standard, with weights determined by: (a) the reliability of the relevant exercise scores, (b) the variability of scores on the exercises, and (c) panel members' ratings of the importance of the exercises in assessing candidates' performances on the content standard. Other models for deriving performance standards on a content standard will also be considered, including multihurdle, conjunctive, and mixed-model approaches.

In the spirit of attempting to determine cost effective approaches to applying the extended Angoff standard-setting method, a modification that uses candidates as judges will be considered. After they have completed all of the assessment center exercises, candidates for National Board Certification will be asked (during a debriefing session) to provide two types of judgments: 1) What is the minimum score they think should be allowed for a teacher to be classified as a highly accomplished teacher on each of the exercises that assesses a particular content standard? This question will be posed by using a sampling plan that will yield sufficient information across candidates for all content standards, and 2) What are the absolute and relative importance of the various exercises that assess candidates' performances on the content standard? A comparison will be made between the results based on panel members' judgments and those based on ratings provided by candidates. If candidates' judgments and panelists' judgments are found to result in exchangeable standards, use of candidates as judges will be recommended, since significant economies could be realized.

BEST COPY AVAILABLE

Conclusions

Although the approaches to standard setting described in this paper were considered in the context of the assessment development activities of the National Board for Professional Teaching Standards, their potential application is far more general. Whenever certification, licensure, or selection decisions require the integration of information on candidates' performances on multiple dimensions of assessment, equity considerations require a well-structured, replicable approach to decision making. When assessment strategies employ standardized instrumentation, the requirements of the 1985 *Standards for Educational and Psychological Testing* must be satisfied. Those testing standards require that performance standard-setting procedures be reliable, public, and well documented. The strategies described and contrasted in this paper should, if carefully applied, fully satisfy the requirements of the 1985 *Test Standards* and be widely applicable in a variety of certification and selection contexts.

References

American Psychological Association (1985). *Standards for educational and psychological testing*. Washington, DC: Author.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd. ed.). New York: Macmillan.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*, 137-172.

Ebel, R. L. (1972). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Gardiner, P. C. & Edwards, W. (1975). Public values: Multiattribute utility measurement for social decision making. In M. F. Kaplan & S. Schwartz (Eds.). *Human judgment and decision processes*. New York: Academic Press.

Hobson, C. J., Mendel, R. M. & Gibson, F. W. (1981). Clarifying performance appraisal criteria. *Organizational Behavior and Human Performance, 28*, 164-188.

BEST COPY AVAILABLE

Hobson, C. J. & Gibson, F. W. (1983). Policy capturing as an approach to understanding and improving performance appraisal: A review of the literature. *Academy of Management Review, 8,* 640-649.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4,* 461-475.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd. ed.). New York: American Council on Education/Macmillan.

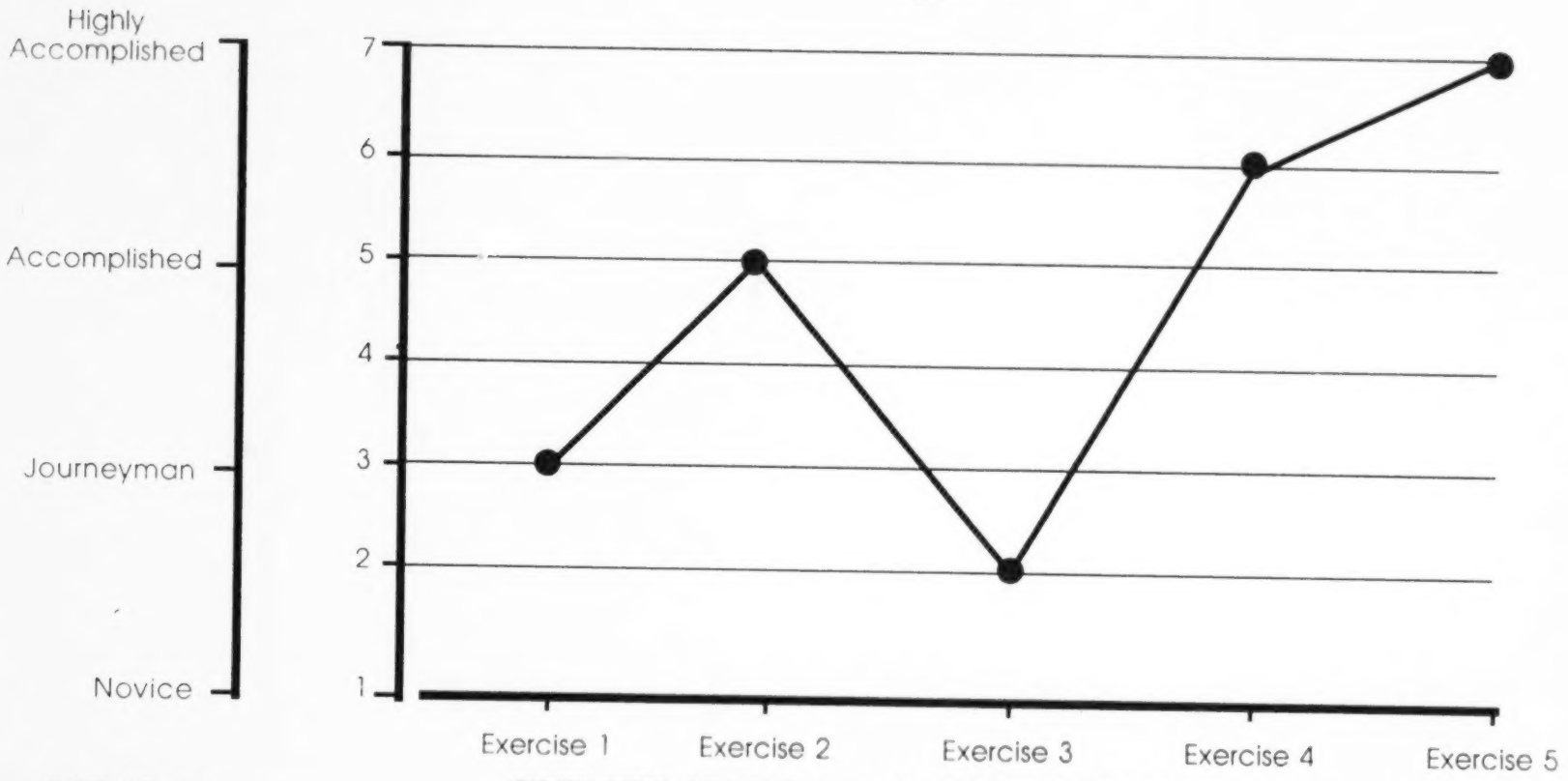
Livingston, S. A. & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance of educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3-19.

Stumpf, S. A. & London, M. (1981). Capturing rater policies in evaluating candidates for promotion. *Academy of Management Journal, 24,* 752-766.

BEST COPY AVAILABLE

Figure 1



RATING ON STANDARD

SIMULATED PROFILE OF CANDIDATE PERFORMANCES ON EXERCISES THAT ASSESS A SINGLE CONTENT STANDARD

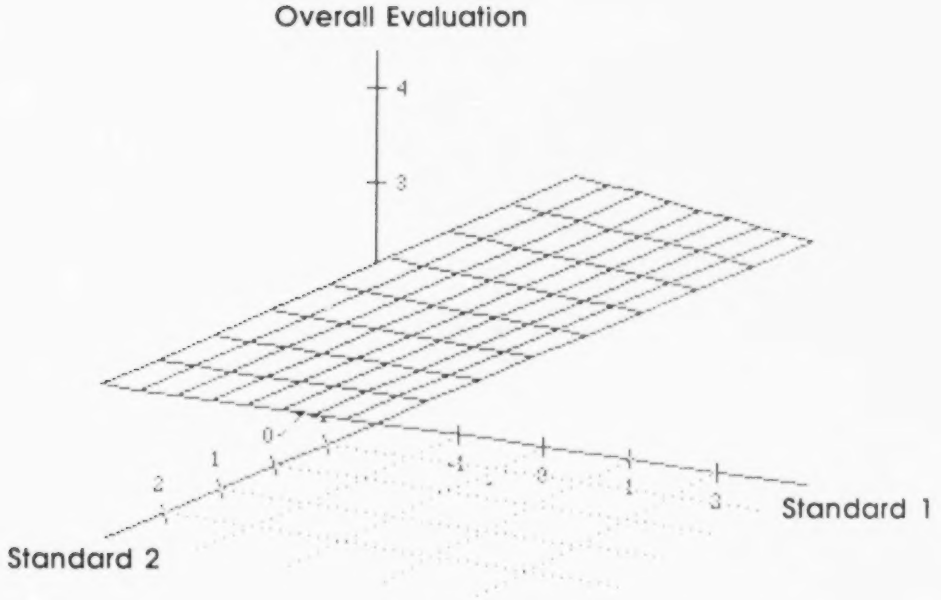


Figure 2

A Compensatory Model for Evaluation of Teaching Performance

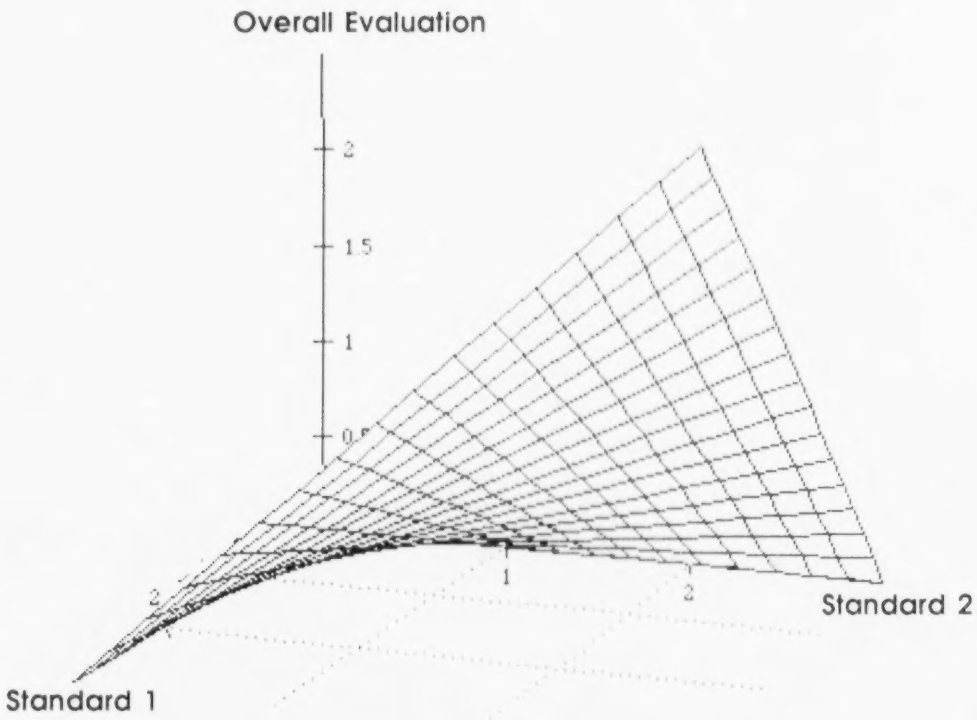


Figure 3

A Conjunctive Model for Evaluation of Teaching Performance

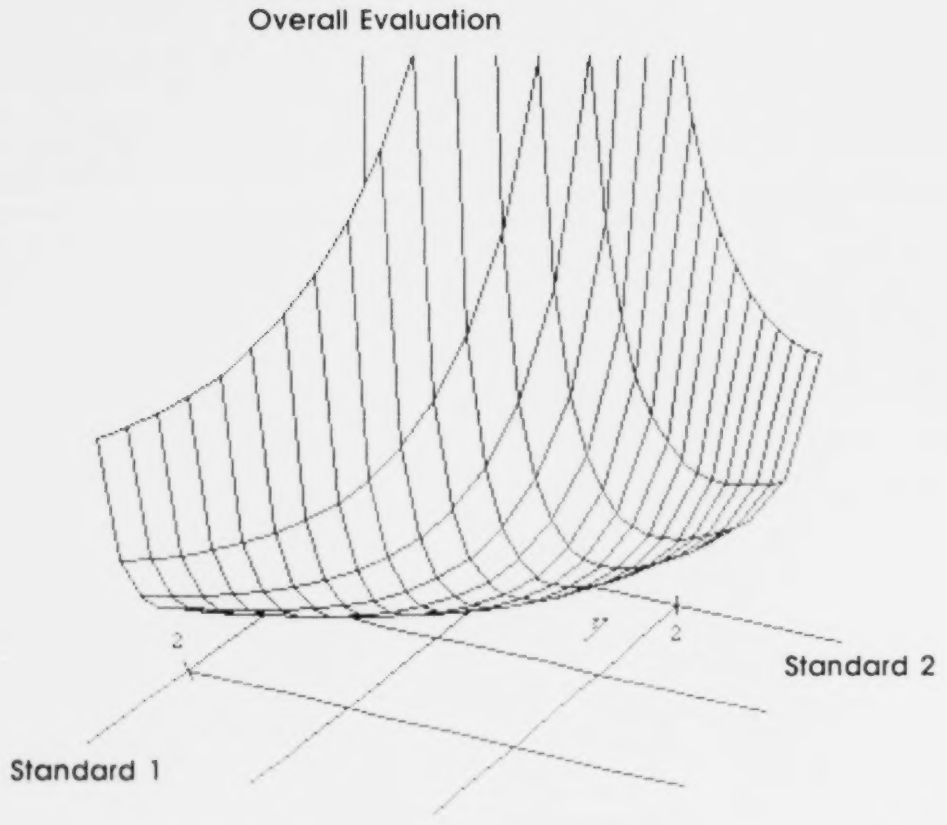


Figure 4

A Disjunctive Model for Evaluation of Teaching Performance